MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2012) **12**, 846–855 doi: 10.1111/j.1755-0998.2012.03153.x

Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics

ARNAUD ESTOUP,* ERIC LOMBAERT,† JEAN-MICHEL MARIN,‡ THOMAS GUILLEMAUD,† PIERRE PUDLO,*‡ CHRISTIAN P. ROBERT§¶** and JEAN-MARIE CORNUET*

**Inra, UMR1062 Cbgp, Montpellier, France, †Equipe 'Biologie des Populations en Interaction', UMR 1301 IBSV INRA-CNRS-Université de Nice-Sophia Antipolis, Sophia-Antipolis, France, ‡I3M, UMR CNRS 5149, Université Montpellier 2, Montpellier, France, §Université Paris Dauphine, CEREMADE, Paris, France, ¶Institut Universitaire de France, Paris, France, **CREST, Paris, France*

Abstract

Comparison of demo-genetic models using Approximate Bayesian Computation (ABC) is an active research field. Although large numbers of populations and models (i.e. scenarios) can be analysed with ABC using molecular data obtained from various marker types, methodological and computational issues arise when these numbers become too large. Moreover, Robert *et al.* (*Proceedings of the National Academy of Sciences of the United States of America*, 2011, 108, 15112) have shown that the conclusions drawn on ABC model comparison cannot be trusted *per se* and required additional simulation analyses. Monte Carlo inferential techniques to empirically evaluate confidence in scenario choice are very time-consuming, however, when the numbers of summary statistics (Ss) and scenarios are large. We here describe a methodological innovation to process efficient ABC scenario probability computation using linear discriminant analysis (LDA) on Ss before computing logistic regression. We used simulated pseudo-observed data sets (*pods*) to assess the main features of the method (precision and computation time) in comparison with traditional probability estimation using raw (i.e. not LDA transformed) Ss. We also illustrate the method on real microsatellite data sets produced to make inferences about the invasion routes of the coccinelid *Harmonia axyridis*. We found that scenario probabilities computed from LDA-transformed and raw Ss were strongly correlated. Type I and II errors were similar for both methods. The faster probability computation that we observed (speed gain around a factor of 100 for LDA-transformed Ss) substantially increases the ability of ABC practitioners to analyse large numbers of *pods* and hence provides a manageable way to empirically evaluate the power available to discriminate among a large set of complex scenarios.

Keywords: Approximate Bayesian Computation, coalescence, discriminant analysis, evolutionary scenario, model probability, molecular markers, population genetics

Received 13 January 2012; revision received 28 March 2012; accepted 2 April 2012

Introduction

One prospect of current biology is that molecular data will help us to reveal the complex demographic processes that have acted on natural populations. The extensive availability of various molecular markers and increased computer power have promoted the development of inferential methods and associated softwares (e.g. Beaumont & Rannala 2004; Excoffier & Heckel 2006). Among these novel methods, Approximate Bayesian Computation method (ABC; Beaumont *et al.* 2002) is increasingly used to make inferences from large data sets for complex models in population and evolutionary biology (e.g. Estoup *et al.* 2004; Jakobsson *et al.* 2006; Fagundes *et al.* 2007; Rosenblum *et al.* 2007; Neuenschwander *et al.* 2008; Toni *et al.* 2009; Verdu *et al.* 2009; Bazin *et al.* 2010; Estoup & Guillemaud 2010; Ascunce *et al.* 2011). The use of ABC techniques has also been envisaged and successfully processed in other research fields, such as infectious disease epidemiology (e.g. Luciania *et al.* 2009) and systems biology (e.g. Ratmann *et al.* 2009).

General statistical features, practical aspects, and applications of ABC in evolutionary biology have been reviewed in at least three recent papers (Beaumont 2010; Bertorelle *et al.* 2010; Csilléry *et al.* 2010). Briefly, ABC constitutes a recent approach to carrying out model-based inference in a Bayesian setting in which model

Correspondence: Arnaud Estoup, Fax: +33-(0)4-99-62-33-45; E-mail: estoup@supagro.inra.fr

likelihoods are difficult to calculate (owing to the complexity of the models considered) and must be estimated by massive simulations. In ABC, the posterior probabilities of different models and/or the posterior distributions of the demographic parameters under a given model are determined by measuring the similarity between the observed data set (i.e. the target) and a large number of simulated data sets; all raw data sets (i.e. multilocus genotypes or individual sequences) are summarized by so called summary statistics (Ss). Examples of such Ss in population genetics are the mean number of alleles or heterozygosity per population and F_{ST} or genetic distances between pairs of populations. In practice, ABC users can base their analysis on simulation programs and then use statistical software to postprocess their simulation outputs. Several ABC programs have recently been developed to provide nonspecialist users with integrated solutions. They vary in the extent to which they are user friendly and they can be used for both data simulation and some postprocessing steps (see Table 1 in Bertorelle *et al.* 2010).

Although the methodology presented here is of more general interest, the present work focuses on population genetics applications and applies to the model choice question. In this context, models are evolutionary scenarios for which relative supports are compared through their posterior probabilities. Choosing among a finite set of scenarios is crucial when doing inferences about evolutionary history and processes for at least two reasons: (i) it allows making general conclusions about major evolutionary events (e.g. admixture between populations, occurrence of bottleneck events or identification of source populations) and (ii) it makes it possible to estimate posterior probabilities of parameters assuming a single scenario if the later is strongly supported (see the reviews of Bertorelle *et al.* 2010; Csilléry *et al.* 2010; Estoup & Guillemaud 2010 for various illustrations regarding model choice). When processing ABC analyses, all the models are generally simulated the same number of times. This is equivalent to giving the same prior probability to each model under comparison and zero probability to any other model. In the final set of retained simulations (those that have Ss close to the target's), the data sets produced by the more supported models will be overrepresented, and the data sets produced by other models will be under-represented or even absent. Intuitively, the probability of a model is related to the relative frequency of the data sets it produces, which are among the retained simulations (Weiss & von Haeseler 1998; Pritchard *et al.* 1999). This frequency may be taken as an estimate of the posterior probability of a model, but this estimate is rarely accurate in complex models when, inevitably, they are too few retained simulations or they also contain data sets not closely matching the observed data (e.g. Guillemaud *et al.* 2010). Recently, Leuenberger & Wegmann (2010) proposed the use of a parametric general linear model to adjust the model frequencies in the retained simulations. However, the most used and tested method, also available in integrated ABC packages such as DIYABC (Cornuet *et al.* 2008, 2010), is the adjustment based on the polychotomous logistic regression introduced by Beaumont (2008) (see also Fagundes *et al.* 2007; Cornuet *et al.* 2008). The coefficients for the regression between a model indicator (response) variable and the simulated Ss (the explanatory variables) can be estimated, allowing the estimation of the posterior probability for each model at the intercept condition where observed and simulated Ss coincide. Confidence intervals (i.e. 95% CI) of the probabilities can be computed as suggested by Cornuet *et al.* (2008).

Large numbers of populations and loci can be analysed with ABC, and there is no limit to the number and complexity of the models (hereafter named scenarios) considered. However, several issues arise when the number of populations becomes too large. The number of Ss to be manipulated increases considerably with the number of populations. This is especially true when different types of markers requiring different types of Ss are considered in the same analysis. A too large number of Ss may be of concern because ABC algorithms attempt to sample from a small multidimensional sphere around the observed statistics. The more Ss, the more difficult it becomes to match the observations closely and increasing the number of simulations may not be sufficient to deal with this issue (Beaumont *et al.* 2002). This phenomenon, which may potentially degrade the estimations of posterior distributions of demo-genetic parameter as well as those of model posterior probabilities, is often referred to as the 'curse of dimensionality' (e.g. Beaumont *et al.* 2002; Blum & François 2009). There may be also a problem of colinearity among explanatory variables (Ss) resulting in instability of the regression when (too) many Ss are introduced (Besley *et al.* 2004; Bazin *et al.* 2010). Recent improvements of ABC get round these problems using dimension reduction techniques, including a nonlinear feed-forward neural network (Blum & François 2009) and partial least squares (PLS) regression (Wegmann *et al.* 2009; see also Bazin *et al.* 2010). At least some algorithms of this type have been implemented in the package ABC-toolbox (Wegmann *et al.* 2010). The added value of such algorithms in the context of complex models and large data sets remains, however, to be thoroughly tested (Bertorelle *et al.* 2010). Most importantly, although the model itself can be considered as an additional parameter to infer, the PLS dimension reduction technique applies to a continuous response variable. Therefore, this technique can be applied to the estimation of posterior distributions of demographic and genetic parameters under a given model and not to the computation of posterior probabilities of models, the latter corresponding to a discrete response variable. Initially developed for the estimation of posterior distributions of demographic and genetic parameters, neural networks might theoretically be applied to model choice (Ripley 1996), but, to our knowledge, this has not been tested and achieved in practice, at least in the context of complex models and large data sets.

Robert *et al.* (2011) have shown that, because ABC algorithms involve an unknown loss of information induced by the use of insufficient summary statistics, the conclusions drawn on model comparison cannot be trusted *per se* and required further simulation analyses. As pointed by Bertorelle *et al.* (2010) and Robert *et al.* (2011) among others, confidence in model choice may be nevertheless empirically evaluated by processing Monte Carlo evaluation of false allocation rates (type I and II errors) based on ABC posterior probabilities computed from simulated pseudo-observed data sets. A version of this exploratory analysis is already provided in the DIYABC software (Cornuet *et al.* 2008, 2010). This evaluation, based on the simulation and analysis of pseudo-observed data sets (hereafter named *pods*), represents a useful and manageable quality assessment for practitioners but is very time-consuming. The polychotomous logistic regression used to estimate scenario probabilities requires the computation of a matrix involving a very large number of loops [i.e. (number of compared scenarios)^2 × (number of Ss)^2 × (number of selected simulated data sets close to the target data set)] at each iteration of the Newton–Raphson method (Cornuet *et al.* 2008). This makes computation particularly time-consuming when the number of scenarios and Ss become large. Moreover, computations involve several large matrices and probabilities that are sometimes simply not computable when the computer memory space is not large enough. This is of particular concern when type I and type II errors have to be computed from a large number of *pods*. As previously stressed, such computations are nevertheless more and more requested by ABC experts for assessing the power to discriminate among scenarios (e.g. Fagundes *et al.* 2007; Verdu *et al.* 2009; Bertorelle *et al.* 2010; Lombaert *et al.* 2010; Robert *et al.* 2011).

In this paper, we describe a methodological innovation to more efficiently process ABC scenario probability estimation using linear discriminant analysis (LDA) transformations on Ss before computing the logistic regression. We first describe the principle and goals of the method. We then use simulated *pods* to assess its main features (precision and computation time) in comparison with probability estimation using logistic regression on raw (i.e. not LDA transformed) Ss. Finally, we illustrate the method on real microsatellite data sets produced by Lombaert *et al.* (2011) to make inferences about the worldwide routes of invasion of the coccinelid *Harmonia axyridis*.

Materials and methods

Linear discriminant analysis

The LDA is a standard technique for supervised classification. For a modern and comprehensive presentation of LDA, we invite readers to refer to classical textbooks such as Ripley (1996), McLachlan (2004) or Hastie *et al.* (2009). The LDA dates back to Fisher (1936) who proposed the dimension reduction technique that contributed to the popularity of LDA. Actually, the classifier estimated with the LDA depends only on some linear projection of the data set onto a linear subspace whose dimension is smaller than the number of groups, denoted by K. It is not our purpose here to explain how this low-dimensional projection of the data can further lead to a LDA classifier that provides automatic rules to classify a new data point to the class with the largest posterior probability. As a matter of fact, we are here only interested in the dimension reduction part of LDA and hence in the construction of the $(K - 1)$ discriminant variables. Those discriminant variables are noncorrelated, linear combinations of the original variables that maximize the between-class variance relative to the within-class variance, which is assumed identical among the different classes. This minimizes the overlap between the classes when projected on the discriminant subspace if the within-class distribution were Gaussian. Note that the discriminant variables are ordered with respect to their ability to move the classes further apart.

In the methodological framework considered here (i.e. that of computing posterior probabilities of scenarios using ABC), we used LDA to transform the set of usually large number J of summary statistics (Ss) into $(K - 1)$ independent variables maximizing the differences among the K compared scenarios (assuming $K < J$). The goal was to reduce the dimension of the set of explanatory variables from J nonindependent to $(K - 1)$ independent variables, whatever the value of J. Certainly, variance of the Ss varies among the different scenarios. Even in that case, however, the projection onto the discriminant subspace was proved relevant as a dimension reduction technique; see the classical textbooks cited previously. It is worth noting that we also weighted the simulated data sets to give more importance to the ones that are closer to the observed data set. The LDA functions were used to transform both the (raw) simulated and observed Ss. Details on LDA computations and transformation of Ss are given in the Appendix S1 (Supporting information).

We first recapitulate how computation of the discriminant variables was included in practice as a single additional step of the ABC process to allow the computation of the posterior probabilities of scenarios.

Step 1: We selected a subset of x% (typically 1%) best simulations in a standard ABC reference table (i.e. the table where parameter values drawn from priors and corresponding simulated Ss have been recorded) usually including 10^6 simulations for each of the K compared scenarios. This selection was based on the standard normalized Euclidian distance computed between the observed and simulated 'raw' (i.e. not transformed) Ss (e.g. Beaumont *et al.* 2002) and hence corresponded to the x% simulations with the smallest Euclidian distances.

Step 2 (LDA step; see Appendix S1, Supporting information for details): we used LDA to transform the raw Ss of this subset of x% best simulations into (K − 1) discriminant variables maximizing the differences among the K compared scenarios. When computing LDA functions, we weighted the simulated data sets with the Epanechnikov kernel commonly used in the local regression (equation 5 in Beaumont *et al.* 2002).

Step 3: We estimated the posterior probabilities of each competing scenario by polychotomous logistic regression (Cornuet *et al.* 2008) on the x% best simulated data sets now summarized by (K − 1) discriminant variables instead of J nonindependent variables (i.e. raw Ss statistics). Confidence intervals (i.e. 95% CI) were computed for each posterior probability using the (K − 1) independent variables following Cornuet *et al.* (2008).

Hence, our proposal included only a single additional step (i.e. Step 2) when compared to the computation traditionally proposed by different authors (e.g. Fagundes *et al.* 2007; Beaumont 2008; Cornuet *et al.* 2008, 2010). Processing Step 2 substantially decreases the number of explanatory variables through the production of LDA variables maximizing the differences among the compared scenarios. This provides three main advantages. First, computation of scenario probabilities using the polychotomous regression of Step 3 becomes (much) faster and sometimes simply feasible. Second, a lower number of explanatory variables may also improve the accuracy of the ABC approximation, particularly when the number of simulations is not large enough to offset the number of Ss. Finally, using LDA-transformed Ss avoids correlations among explanatory variables.

Tests on simulated data sets

Pseudo-observed data sets (*pods*) were simulated from a set of known scenarios and prior distributions to compare posterior probabilities obtained through the logistic regression performed on both LDA-transformed and raw Ss. The *pods* were defined to mimic the real microsatellite data set of the ABC analysis 1 processed by Lombaert *et al.* (2011) on the invasive coccinelid *Harmonia axyridis*. The *pods* hence included 18 microsatellites genotyped in five population samples (18–35 individuals per population samples). This data set was produced to make inferences about the origin of the invasive *H. axyridis* population established in eastern North America in 1988 (ENA), considering altogether two populations from the native range, two strains used for biocontrol release and one (target) population from the introduction range (ENA). In this analysis, Lombaert *et al.* (2011) defined ten competing scenarios considering a native or biocontrol population as a source for ENA or admixture between them (see Lombaert *et al.* 2010, 2011 for details).

As in analysis 1 of Lombaert *et al.* (2011), genetic variation within and between populations was summarized in the *pods* using a set of (raw) statistics traditionally employed in ABC (Cornuet *et al.* 2008, 2010; Guillemaud *et al.* 2010). For each population and each population pair, we used the mean number of alleles per locus, the mean expected heterozygosity and the mean allelic size variance. The other statistics used were the mean ratio of the number of alleles over the range of allele sizes, pairwise F_{ST} values, mean individual assignment likelihoods of population i assigned to population j and the maximum likelihood estimate of admixture proportion. The total number of Ss was 86.

We choose this particular scenarios-priors-Ss setting because it had the potential to fairly illustrate our new methodological developments based on LDA-transformed Ss. This setting was characterized by relatively high (mean) type I error rates (*c.* 0.40, owing to the large prior parameter space used to generate *pods*, this space included 'areas' for which the discrimination among scenarios was difficult) and relatively small (mean) type II error rates (*c.* 0.07). High type I error rates correspond to situations where probability values of the target scenario can be small to high depending on the parameter values of the analysed *pod*, hence virtually including the entire spectrum of probabilities between 0 and 1. This allows a better (and fairer) comparison of results between raw and LDA-transformed Ss (cf. it is difficult to compare probability estimations when all values are between say 0.95 and 1.0). Moreover, this particular setting was chosen because it corresponded to complex evolutionary models and large data sets that nevertheless could be analysed for a large number of *pods* using logistic regression on both LDA-transformed and raw Ss. More complex data and scenario settings (with larger number of scenarios and/or raw Ss) were computationally too demanding to obtain probability estimations on a large enough number of *pods* in a manageable time using logistic regression on raw Ss (i.e. <15 min per *pod* on a single standard biprocessor computer). The results presented here were, however, qualitatively similar to those obtained considering various alternative settings (with smaller or larger numbers of scenarios and/or raw Ss) that we have also tested with our methodological innovation (results not shown).

The ABC analyses of the *pods* were performed using parameter values drawn from the prior distributions described in Table S1 (Supporting information) and by simulating 10^6 data sets for each of the ten competing scenarios. For each *pod*, we estimated the posterior probabilities of the scenarios using a polychotomous logistic regression on the 1% of simulated data sets closest to the observed data set, considering either LDA-transformed or raw Ss.

We produced a first set of 500 *pods* under scenario 5 (the scenario selected after ABC treatment by Lombaert *et al.* 2011), drawing parameters values into the distributions described in Table S1 (Supporting information). This scenario 5 is presented graphically in Fig. S1 (Supporting information); the nine other competing scenarios correspond to alternative source(s) of the target-introduced population (see Lombaert *et al.* 2011 for details). For each *pod*, we used the logistic regression on either the 9 LDA-transformed or the 86 raw Ss to estimate the posterior probability and 95% CI of scenario 5 relatively to the set of ten compared scenarios. The number of iterations of the Newton–Raphson algorithm used by the logistic regression computations and the mean time of each iteration were also recorded for each *pod*.

We then produced a second set of 1000 *pods* including 10 subsets of 100 *pods* simulated under each of the 10 compared scenarios, drawing parameter values from the same distributions (Table S1, Supporting information). Each *pod's* subset was used to estimate type I and type II errors on scenario choice using either the nine LDA-transformed or the 86 raw Ss. Type I error of a given scenario is the proportion of *pods* simulated from this scenario for which this scenario does not have the highest posterior probability. Type II error is the proportion of *pods* for which the scenario with the highest posterior probability is not the given true one.

Finally, we evaluated the impact of the dimensionality of the simulated data set (i.e. the 'curse of dimensionality' mentioned in the Introduction section), using either the nine LDA-transformed or the 86 raw Ss. For different amounts of simulated data sets, we estimated the type I and II error rates from 500 *pods* simulated under scenario 5 (type I error for scenario 5) and 500 *pods* simulated under scenario 1 (type II error for scenario 5 which in this case corresponds to the proportion of times that scenario 5 was selected when *pods* have been produced under scenario 1). Scenario 1 was chosen to evaluate type II errors because this scenario has shown the largest type II errors in the above-mentioned analyses. To consider different dimensionalities of simulated data sets, we decreased the number of data sets simulated for each of the ten compared scenarios from 10^6 to 10^4, keeping the proportions of data sets closest to the observed data set selected for the logistic regression at 1% of the total number of simulated data sets.

All analyses were processed on a 2 CPU Intel Xeon X5472 computer (Windows XP platform, 32 bits system, 4 Gb of RAM) using a modified version of the package DIYABC V1. This modified version is available under request from AE. LDA transformation of Ss before logistic regression will be implemented in a new multiplatform version of DIYABC that will be freely available later in 2012.

Tests on real data sets

We used the real microsatellite data sets of Lombaert *et al.* (2011) to compare scenario choice and probability estimation computing logistic regression on both LDA-transformed and raw Ss. These data sets, which included 18 microsatellites genotyped on five to eight population samples (18–42 individuals per population samples), were used to make five consecutive ABC analyses about the worldwide routes of invasion of the coccinelid *H. axyridis*, considering altogether populations from the native range, the introduction range and biocontrol release actions, with potential admixture between them (see Lombaert *et al.* 2010, 2011 for details).

We used prior distributions and Ss identical to those described in the previous section (Tests on simulated data sets; Table S1, Supporting information). Following Lombaert *et al.* (2010, 2011), we performed five consecutive ABC analyses of invasion scenarios involving successive *H. axyridis* outbreaks that were successively recorded in the invaded range. As previously detailed, analysis 1 dealt with the introduction pathway for the first recorded outbreak in eastern North America in 1988, defining 10 competing scenarios. Analysis 2 dealt with the second outbreak recorded in western North America in 1991, taking into account the scenario selected in analysis 1, hence defining 15 competing scenarios. The European and South American outbreaks in 2001 were addressed in analyses 3 and 4, respectively (15 scenarios for each outbreak), taking into account the scenario selected in analysis 1 and 2. Finally, the African outbreak in 2004 was considered in analysis 5 (28 scenarios), taking into account the scenarios selected in analyses 1, 2, 3 and 4. The total number of raw Ss varied from 86 (analysis 1) to 223 (analysis 5), whereas the total number of LDA-transformed Ss varied from 9 (analysis 1) to 27 (analysis 5).

The ABC analyses were performed by simulating 10^6 microsatellite data sets for each competing scenario in the first four analyses and 5×10^5 data sets per scenario in analysis 5 because of the high number of scenarios (28) and raw summary statistics (223) which made a larger analysis computationally too heavy, even when using LDA-transformed Ss. For each of the five analyses, we estimated the posterior probabilities of the competing scenarios using a polychotomous logistic regression on the 1% of simulated data sets closest to the observed data set, considering either LDA-transformed or raw Ss. Computation times were also recorded to illustrate the gain obtained in computation speed when using LDA-transformed Ss.

Finally, we evaluated the impact of the number of simulated data sets recorded in the reference table for analysis 1 on the estimation of the probability of scenario 5 using either LDA-transformed or raw Ss. To this aim, we decreased the number of data sets simulated for each of the ten compared scenarios from 10^6 to 10^4, keeping the proportions of data sets closest to the observed data set selected for the logistic regression at 1% of the total number of simulated data sets.

All analyses were processed on a 2 CPU Intel Xeon E5540 computer (Windows XP platform, 32 bits system, 4 Gb of RAM) using a modified version of the package DIYABC V1 (available under request from AE).

Results

Tests on simulated data sets

Figure 1A illustrates the strong correlation between the probability values of scenario 5 obtained from *pods* computing logistic regression on LDA-transformed Ss and raw Ss (Pearson's correlation coefficient = 0.940). One can see, however, a trend for a globally slightly lower scenario probability with LDA-transformed Ss (see linear regression equation in the legend of Fig. 1A). Figure 1B shows that 95% CI are almost always smaller with LDA-transformed Ss.

Figure 2 summarizes the type I and II error rates obtained with LDA-transformed and raw Ss. We found that these error rates substantially varied among scenarios but were to a large extent similar for both methods for a given scenario. *P*-values computed using Fisher's exact test were higher than 0.6 for all scenarios for mean type II errors and were lower than 5% for a single scenario for type I errors (P = 0.047 for scenario 7; *P*-value nonsignificant after applying the false discovery rate correction method of Benjamini & Hochberg 1995).

The gain in computation time with LDA-transformed Ss was high. First, the number of iterations needed to reach convergence during the logistic regression analysis was lower with LDA-transformed Ss (mean = 7.320, SD = 1.420) than with raw Ss (mean = 9.190, SD = 2.250). Second, the mean time of each such iteration was considerably smaller with LDA-transformed Ss (mean = 7.034 s, SD = 0.791) than with raw Ss (mean = 888.146 s, SD = 65.374). This translated into a computation speed increase by a mean factor 128.128 (SD = 19.482) per iteration and 163.601 (SD = 46.456) for a completed logistic regression analysis. The computation time for the LDA transformation of raw Ss before the regression was negligible.

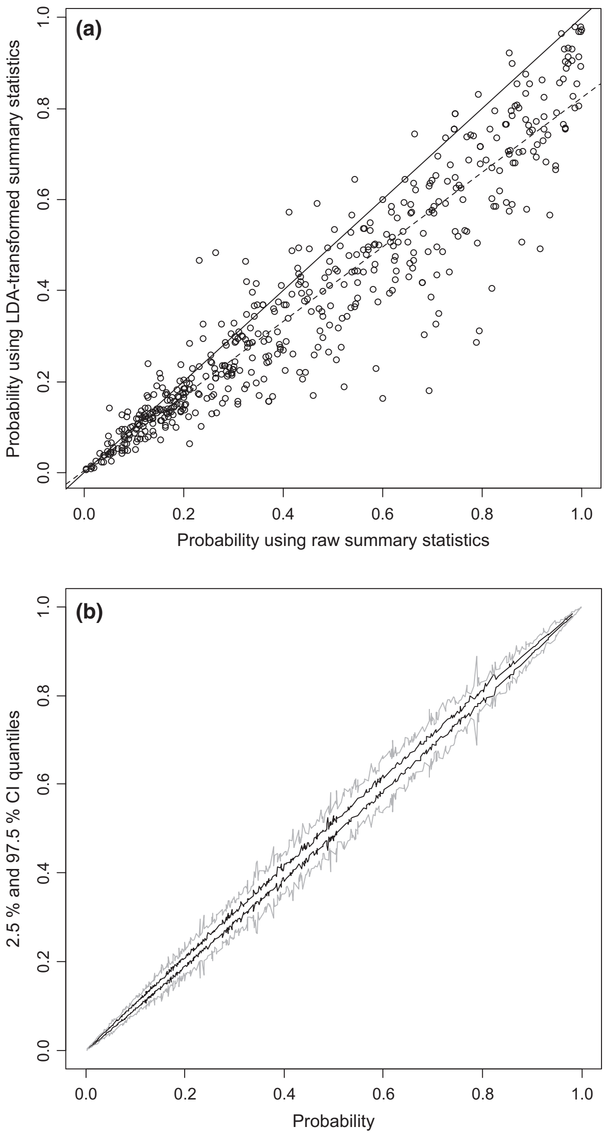

Fig. 1 Probability estimations of scenario 5 computed using linear discriminant analysis (LDA)-transformed or raw summary statistics for 500 pods simulated under scenario 5 (10 scenarios compared). (a) Pearson's correlation coefficient between probability estimations = 0.940 (95% CI = [0.928, 0.949]). Solid line: $y = x$; dotted line: linear regression line $y = 0.818436x + 0.004878$. (b) 95% CIs (i.e. 2.5% and 97.5% quantiles) for each probability values obtained from either LDA-transformed summary statistics (black lines) or raw summary statistics (grey lines).

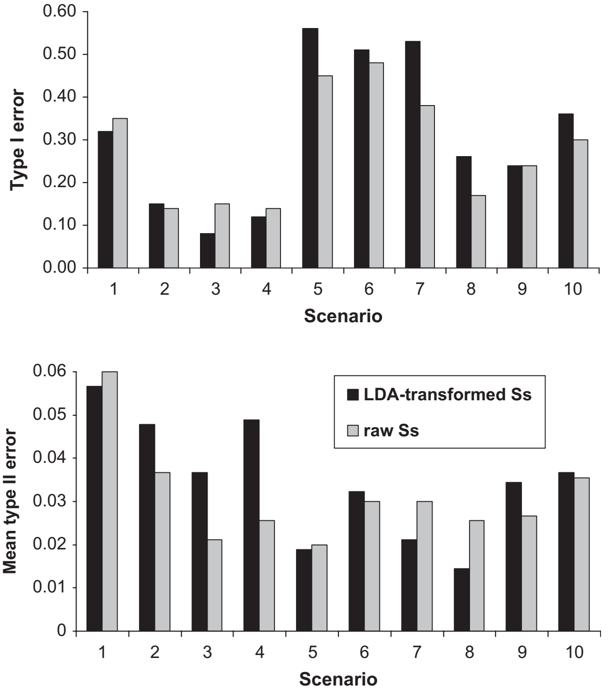

Fig. 2 Confidence in discriminating scenarios using linear discriminant analysis-transformed or raw summary statistics.Note: Type I error: exclude scenario *x* when it is actually scenario *x*. Type II error: choose scenario *x* when it is not scenario *x*. Results are based on 100 *pods* per scenario (total of 10 compared scenarios). The compared scenarios correspond to variants of the scenario 5, the latter being detailed in Fig. S1 (Supporting information).

Table 1 Type I and II error rates estimated for different numbers of simulated data sets

	Number of simulated data sets for each of the 10 compared scenarios				
	10^6	10^5	5×10^4	2×10^4	10^4
Type I error					
LDA-transformed Ss	0.560	0.556	0.584	0.592	0.622
Raw Ss	0.450	0.492	0.530	0.536	0.624
Type II error					
LDA-transformed Ss	0.056	0.056	0.052	0.062	0.080
Raw Ss	0.060	0.062	0.072	0.088	0.116

Type I error rates were estimated for scenario 5 from 500 *pods*. Type II errors were estimated for scenario 5 when simulating 500 *pods* under scenario 1. The number of data sets simulated for each of the 10 compared scenarios decreased from 10^6 to 10^4, keeping the proportions of data sets closest to the observed data set selected for the logistic regression at 1% of the total number of simulated data sets.
LDA, linear discriminant analysis.

Results summarized in Table 1 indicate that we did not face the curse of dimensionality problem (see definition in the Introduction section) at least in the present setting. Even for a large number of Ss and a strongly degraded number of simulated data sets including only 10^4 data sets per scenario (total of 10^5 data sets for the ten compared scenarios in this case), the error rates did not dramatically increase. The increase in type I and II error rates with smaller data sets is (only) slightly faster for raw Ss than for LDA-transformed Ss.

Tests on real data sets

As will be further illustrated later on real data sets, our methodological innovation is particularly attractive when practitioners have to deal with a large number of complex scenarios involving a large number of Ss. Table 2 summarizes our results on scenario choice and probability estimation computing logistic regression on both LDA-transformed and raw Ss obtained on the real microsatellite data sets of Lombaert *et al.* (2011). For each of the five consecutive analyses, the same scenario had the highest probability and was hence selected using either LDA-transformed or raw Ss. The probabilities of the most likely scenarios were slightly smaller with LDA-transformed Ss for analyses 1, 3 and 4, and slightly larger for analysis 2. In contrast to computation based on LDA-transformed Ss, analysis 5 could not be processed with raw Ss owing to computer memory overflow. In all analyses, the 95% CI of the most likely scenario never overlapped those of competing scenarios. As found with simulated *pods*, 95% CI with LDA-transformed Ss were smaller than those with raw Ss.

In agreement with *pods* analyses, the gain in computation time with LDA-transformed Ss was substantial. For all analyses, both the number of iterations needed to reach convergence during the logistic regression and the mean computation time for each such iteration were smaller with LDA-transformed Ss. This translated into a computation speed increase by a factor 72–101 per iteration and 93–159 for a completed logistic regression analysis.

Figure 3 indicates that analysis 1, processed either on LDA-transformed or raw Ss, is rather robust to the potential difficulties associated with the curse of dimensionality. Estimations of the probability of scenario 5 start to fluctuate substantially and 95% CIs to increase considerably for simulation efforts including <2×10^5 data sets per scenarios. No obvious differences could be observed between LDA-transformed and raw Ss.

Discussion

Model comparison is an active research field among the widespread developments currently undergone in ABC (e.g. Beaumont *et al.* 2009; Beaumont 2010; Bertorelle *et al.* 2010; Csilléry *et al.* 2010; Robert *et al.* 2011). Here, we propose a methodological innovation to deal with the

Table 2 Scenario choice and posterior probability estimated from either LDA-transformed or raw summary statistics when considering the real microsatellite data sets of Lombaert *et al.* (2011)

	Logistic regression on raw summary statistics					Logistic regression on LDA-transformed summary statistics					Speed gain
Consecutive ABC analyses (number of simulations per scenario\number of scenarios)	Number of stats	Selected scenario #	Posterior probability [95% CI]	Mean time per NR iteration	Number of NR iterations	Number of stats	Selected scenario #	Posterior probability [95% CI]	Mean time per NR iteration	Number of NR iterations	Per iteration\over all iterations
Analysis 1 (10^6\10 scenarios)	86	5	0.6242 [0.5767, 0.6717]	5′ 05″	11	9	5	0.5420 [0.5325, 0.5516]	3″	7	101.7\159.8
Analysis 2 (10^6\15 scenarios)	124	1	0.4425 [0.3746, 0.5105]	38′ 45″	11	14	1	0.5767 [0.5559, 0.5976]	31″	7	75.0\117.9
Analysis 3 (10^6\15 scenarios)	124	13	0.8134 [0.7107, 0.9160]	38′ 38″	9	14	13	0.7487 [0.7214, 0.7760]	32″	6	72.4\93.1
Analysis 4 (10^6\15 scenarios)	124	4	0.9489 [0.9315, 0.9663]	33′ 36″	9	14	4	0.9227 [0.9139, 0.9315]	27″	7	74.7\96.0
Analysis 5 (5×10^5\28 scenarios)	223	NC*	NC	NC	NC	27	4	0.6864 [0.6456, 0.7272]	6′ 11″	7	NC

The probabilities of the competing scenarios were computed using a logistic regression on the 1% of simulated data sets closest to the real *Harmonia axyridis* data sets.
NR iterations, Newton–Raphton iterations (Cornuet *et al.* 2008); NC, not computable; ABC, Approximate Bayesian Computation; LDA, linear discriminant analysis.
*Because the full computation of analysis 5 was not feasible (Lombaert *et al.* 2011), an alternative method was used to compare scenarios by first setting aside 11 scenarios using the direct approach (Cornuet *et al.* 2008) on the 0.01% data sets closest to the observed data set. The scenario 4 was then selected owing to its highest posterior probability in a subsequent analysis (using polychotomous logistic regression and raw Ss) performed on the 19 remaining scenarios (see Lombaert *et al.* 2011 for details).

discrimination among a large set of complex scenarios through more efficient ABC probability computation using a LDA on Ss before the logistic regression analysis. Statistical methods to select appropriate Ss to optimize model selection are still under development and discussed (see for instance Fearnhead & Prangle 2012 and associated discussions). Our LDA-based transformation of Ss represents a practical and straightforward way to tackle this question.

We show, using both simulated and real data sets, that posterior probabilities of scenarios computed from LDA-transformed and raw Ss are strongly correlated. LDA-transformed Ss tend, however, to provide slightly lower probability values and hence to be somewhat conservative with respect to scenario discrimination. On the other hand, model probabilities estimated from LDA-transformed Ss are characterized by smaller 95% CI. The later feature is expected to decrease the number of inconclusive results if nonoverlapping of CI is taken as a criterion to select a scenario. When scenario selection is made on the basis of the highest probability, type I and II errors were nevertheless similar for both methods. The lower number of LDA variables used for the logistic regression analysis (e.g. nine LDA-transformed Ss vs. 86 raw Ss in the *pods* we analysed) is likely to explain, to a large extent, both the smaller 95% CIs of probability estimates and the smaller number of iterations needed to reach convergence during the regression.

A major practical advantage of using LDA-transformed Ss is that it substantially decreases the dimension of explanatory variables making computation of scenario probability (much) faster and sometimes simply feasible when the available memory space is not large enough to compute the matrix of second partial derivatives of the likelihood (p1 of Supplementary material in Cornuet *et al.* 2008), as in analysis 5 using the real data set of Lombaert *et al.* (2011). This allows larger data scenarios setting to be analysed. It is worth stressing, however, that because LDA transformation only works with the number of Ss and not on the number of parameters of the models, such transformation should not motivate ABC practitioners to over-parameterize their models.

Faster probability computation increases the ability of ABC practitioners to analyse large numbers of *pods* (for instance, using the option 'Evaluate confidence in scenario choice' in the package DIYABC). It hence makes it easier to process a manageable empirical evaluation of the power to discriminate among a given set of scenarios by computing type I and II errors from sufficiently large number of *pods*, especially for large sets of complex scenarios (see e.g. Robert *et al.* 2011 for theoretical arguments in favour of such experimental explorations). Several authors have suggested to use scenario probabilities computed from *pods* to evaluate type I and II errors to

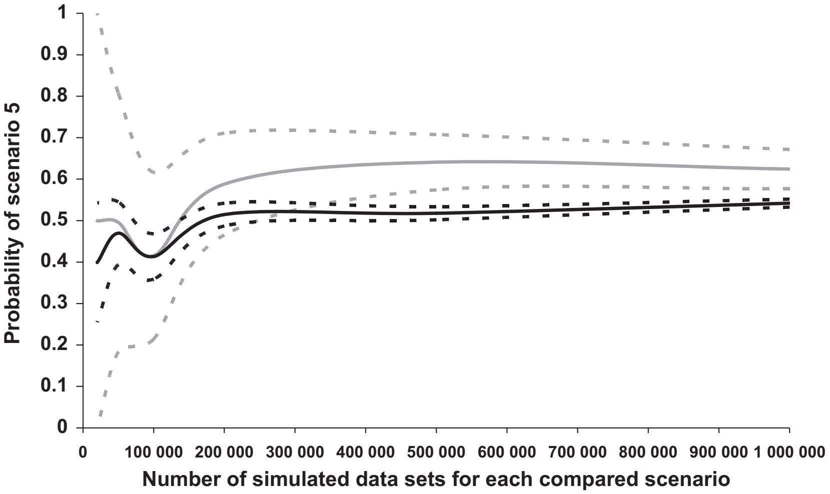

Fig. 3 Probabilities of scenario 5 computed from the real data set of Lombaert *et al.* (2011) for different numbers of simulated data sets.Note: Black = linear discriminant analysis-transformed summary statistics. Grey = raw summary statistics. Plain and dotted lines are for probability estimations and 95% CIs, respectively. Probabilities of scenario 5 were estimated for number of data sets simulated for each of the ten compared scenarios decreasing from 10^6 to 10^4, keeping the proportions of data sets closest to the observed data set selected for the logistic regression at 1% of the total number of simulated data sets.

estimate the posterior probability of a model among a set of *k* models given the observed posterior probability of a real data set, *P* (M_k is the true model | observed estimated posterior probability = *x*). Such computation can then be used to adjust the posterior probabilities estimated from the real data set, taking part of the errors associated with ABC into account (see Fagundes *et al.* 2007).

Other potential advantages of LDA transformation of raw Ss include reducing the difficulties associated with the curse of dimensionality and avoiding correlation among explanatory variables (i.e. multi-colinearity) during the regression step. At least theoretically, the dimensionality issue might be offset by increasing the number of simulations, but the amount of time then needed for concrete implementation might be unreasonable. It is worth stressing, however, that the actual impact of such potential issues remains difficult to assess in a generic manner as it probably differs depending on the analysed observed data set, as well as on the Ss and/or scenario settings. Table 1 and Fig. 3 both indicate a good robustness to the numbers of simulated data sets, as a substantial effect could be observed only for particularly low (and in practice rarely used) number of simulated data sets. Analyses carried on *pods* suggest a slightly better robustness when using LDA-transformed rather than raw Ss, at least when using type I and II error rates as criterion (cf. the slightly smaller increase in errors with smaller data sets for LDA-transformed than raw Ss). It is difficult to know, however, to which extent this result reflects the lower number of LDA variables used for the regression and/or the fact that a substantial number of raw Ss are nonindependent variables.

We believe that our LDA-based methodological innovation will usefully enlarge the tool box available to biologists to make ABC inferences on more complex and hence more realistic demographic processes that have acted on natural populations.

Acknowledgements

This research was financially supported by the French Agence Nationale de la Recherche grants ANR-09-BLAN-0145-01 – EMILE to all authors. We thank the EMILE working group and Eric Bazin for their useful discussions.

References

Ascunce MS, Yang CC, Oakey J *et al.* (2011) Global invasion history of the fire ant *Solenopsis invicta*. *Science*, **331**, 1066–1068.

Bazin E, Dawson KJ, Beaumont MA (2010) Likelihood-Free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics*, **185**, 587–602.

Beaumont MA (2008) Joint determination of topology, divergence time and immigration in population trees. In: *Simulation, Genetics and Human Prehistory* (eds Matsamura S, Forster P & Rrenfrew C), pp. 135–154. McDonald Institute for Archaeological Research, Cambridge, UK.

Beaumont M (2010) Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology and Evolution, and Systematics*, **41**, 379–406.

Beaumont M, Rannala B (2004) The Bayesian revolution in genetics. *Nature Review Genetics*, **5**, 251–261.

Beaumont MA, Zhang WY, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Beaumont MA, Cornuet J-M, Marin J-M, Robert CP (2009) Adaptivity for ABC algorithms: the ABC-PMC scheme. *Biometrika*, **96**, 983–990.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300.

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology*, **19**, 2609–2625.

Besley DA, Kuh E, Welsch RE (2004) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley & Sons, Inc., Hoboken, New Jersey.

Blum MGB, François O (2009) Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, **20**, 63–73.

Cornuet J-M, Santos F, Beaumont MA *et al*. (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, **24**, 2713–2719.

Cornuet J-M, Ravigne V, Estoup A (2010) Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics*, **11**, 401.

Csilléry K, Blum M, Gaggiotti O, François O (2010) Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*, **25**, 410–418.

Estoup A, Guillemaud T (2010) Reconstructing routes of invasion using genetic data: why, how and so what? *Molecular Ecology*, **19**, 4113–4130.

Estoup A, Beaumont M, Sennedot F, Moritz C, Cornuet J-M (2004) Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *Bufo marinus*. *Evolution*, **58**, 2021–2036.

Excoffier L, Heckel G (2006) Computer programs for population genetics data analysis: a survival guide. *Nature Review Genetics*, **7**, 745–758.

Fagundes NJR, Ray N, Beaumont M *et al*. (2007) Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 17614–17619.

Fearnhead P, Prangle D (2012) Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society. Series B (Methodological)*, **74**, 1–28.

Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

Guillemaud T, Beaumont M, Ciosi M, Cornuet J-M, Estoup A (2010) Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity*, **104**, 88–99.

Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, Second Edition Springer Series in Statistics. Springer-Verlag, New York.

Jakobsson M, Hagenblad J, Tavaré S *et al*. (2006) A recent unique origin of the allotetraploid species *Arabidopsis suecica*: evidence from nuclear DNA markers. *Molecular and Biological Evolution*, **23**, 1217–1231.

Leuenberger C, Wegmann D (2010) Bayesian Computation and model selection without likelihoods. *Genetics*, **184**, 243–252.

Lombaert E, Guillemaud T, Cornuet J-M, Malausa T, Facon B, Estoup A (2010) Bridgehead effect in the worldwide invasion of the biocontrol harlequin ladybird. *PLoS ONE*, **5**, e9743.

Lombaert E, Guillemaud T, Thomas C, Lawson Handley L-J, Li J *et al*. (2011) Inferring the origin of populations introduced from a genetically structured native range by approximate Bayesian computation: case study of the invasive ladybird *Harmonia axyridis*. *Molecular Ecology*, **20**, 4654–4670.

Luciania F, Sisson SA, Jiang H, Francis AR, Tanaka MM (2009) The epidemiological fitness cost of drug resistance in *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 14711–14715.

McLachlan GJ (2004) *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Interscience, Hoboken, New Jersey.

Neuenschwander S, Largiader CR, Ray N, Currat M, Vonlanthen P, Excoffier L (2008) Colonization history of the Swiss Rhine basin by the bullhead (*Cottus gobio*): inference under a Bayesian spatially explicit framework. *Molecular Ecology*, **17**, 757–772.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**, 1791–1798.

Ratmann O, Andrieu C, Wiuf C, Richardson S (2009) Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 10576–10581.

Ripley BD (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Robert CP, Cornuet J-M, Marin J-M, Pillai NS (2011) Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 15112–15117.

Rosenblum EB, Hickerson MJ, Moritz C (2007) A multilocus perspective on colonization accompanied by selection and gene flow. *Evolution*, **61**, 2971–2985.

Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, **6**, 187–202.

Verdu P, Austerlitz F, Estoup A *et al*. (2009) Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Current Biology*, **19**, 312–318.

Wegmann D, Leuenberger C, Excoffier L (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, **182**, 1207–1218.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABC-toolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, **11**, 116.

Weiss G, von Haeseler A (1998) Inference of population history using a likelihood approach. *Genetics*, **149**, 1539–1546.

A.E. initiated the project, processed the simulations and wrote the paper. J.-M.C. developed the math and the computer code of the linear discriminant analysis, and co-wrote the paper. E.L., and T.G. analysed the real data sets and co-wrote the paper. J.-M.M., P.P., and C.P.R. checked the math and co-wrote the paper.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Graphical representation of the invasion scenario 5 selected after the ABC analysis 1 processed in Lombaert *et al.* (2011).

Table S1 Prior distributions of demographic, historic and mutation parameters used in ABC analyses attempting to retrace the worldwide routes of invasion of *Harmonia axyridis*.

Appendix S1 Weighted Linear Discriminant Analysis to compare scenarios in DIYABC.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.